The Architecture of Behaviour: A Hierarchical Model of Cognitive Agency and Systemic Failure

1. Introduction: Towards a Unified Computational Taxonomy of Human Action

Current paradigms in psychology and psychiatry frequently suffer from a fragmentation of perspective, relying on trial-and-error interventions that lack a cohesive mechanical foundation. This whitepaper proposes "The Architecture of Behaviour," a vertically integrated computational model that frames human action as a hierarchical stack of processing layers. To understand this architecture, one should look to the **OSI (Open Systems Interconnection) model** in computer networking. Just as the OSI model defines a "Layered Protocol Stack" where the Physical Layer (L1) must function for the Application Layer (L7) to transmit data, the human L-Stack suggests that cognitive agency is contingent upon the integrity of lower-level biological and signal-processing protocols.The core thesis of this framework is that genuine human agency is localized within a specific, metabolically expensive "Explore-Exploit" transition—technically defined as the **Explore-Exploit Gap** . Located neurobiologically in the prefrontal cortex, specifically involving the **frontopolar cortex (FPC)** , this gap between stimulus and response allows the system to temporarily suspend automated scripts to evaluate novel data. Agency, therefore, is not a constant state but a fragile, intermittent capacity susceptible to six distinct, classifiable failure states.This model synthesizes Karl Friston's **Free Energy Principle** (hierarchical predictive coding), Antonio Damasio's **Somatic Marker Hypothesis** (interoceptive guidance), and Bernard Baars' **Global Workspace Theory** (conscious broadcast). It provides a functional, computational heuristic for diagnosing systemic malfunctions, mapping the flow of .information rather than serving as a literal anatomical atlas

2. The L-Stack: A Six-Layer Hierarchical Architecture

The system operates as a protocol stack where information flows bidirectionally. Top-down .predictions attempt to "explain away" bottom-up prediction errors

The L-Stack Summary Table

Layer,Functional Role,Ground Truth / Correlate,Primary Objective
L5: The Field,Environmental Programming,"External systems, algorithms, institutions",Install/reinforce L2 scripts
L4: The Mask,Behavioral Output,"Social performance, active inference",Minimize social prediction error
L3: The Pilot,Executive Agency,Frontopolar / Prefrontal Cortex,Transition from exploit to explore
L2: The Analyst,Prediction Library,"Generative models, schemas, scripts",Pattern matching and rationalization
L1: The Animal,Prediction Error,"Subcortical signals, limbic system",Alert system to discrepancies
L0: The Body,Biological Substrate,"Allostatic load, interoceptive inference",Power the cognitive hierarchy

(L0 - The Biological Substrate (Hardware 2.1

L0 is the physical platform—the "Physical Layer"—upon which all higher-order cognition runs. It is defined by **Allostatic Load** (the cumulative wear and tear on the system) and **Interoceptive Inference** . Biological constraints, including sleep hygiene, glucose availability, and systemic inflammation, set the operational ceiling for the stack. When the metabolic energy budget is depleted, the brain prioritizes survival via "throttling," effectively .removing power from the metabolically expensive L3 layer to preserve L0 homeostasis

L1 & L2 - Prediction Errors and Generative Models 2.2

- **L1 (The Animal):** Emotions are conceptualized as **precision-weighted prediction error signals** . Fear indicates a predicted threat; grief signifies a massive, sustained discrepancy where the internal model predicts the presence of a lost attachment .figure
- **L2 (The Analyst):** This is the system's generative library of scripts and schemas. It functions as the brain's internal model of "how the world works." A significant systemic risk is the **Motivated Reasoning** loop, where an L1 impulse recruits L2 to generate a post-hoc rationalization, creating an illusion of logic for an automated .subcortical reaction

(L3 - The Pilot (Executive Agency 2.3

L3 represents the neural seat of agency, specifically linked to the **Frontopolar Cortex (FPC)** and the dorsolateral prefrontal cortex. It facilitates the **Explore-Exploit Gap** : the transition from **Exploitation** (executing automated, low-cost scripts) to **Exploration** (deliberately evaluating novel options). This transition is computationally and metabolically demanding. It is the only point where the system can genuinely update its generative model .rather than merely distorting sensory input to fit existing L2 scripts

L4 & L5 - Behavioral Output and Environmental Fields 2.4

- **L4 (The Mask):** The active inference channel. It delivers social performance (tone, .posture, syntax) calculated to minimize friction within the social environment
- **L5 (The Field):** The external systems—ranging from family dynamics to social media algorithms—that program L2 scripts. L5 systems are frequently designed to .bypass L3 oversight, training the individual's "Analyst" through repetition and reward

The Taxonomy of Systemic Malfunction: The Six Failure States .3

Psychological distress and behavioral errors are framed here as reliability engineering ."failure states" in the processing cascade

Six Failure States Matrix

Failure State,Computational Mechanism,Phenomenological Experience,Optimal Intervention
L1 Override,Subcortical bypass; L1 drives motor output before L3 engages.,"Explosive reaction; ""Amygdala hijack""; immediate regret.",Somatic signal dampening; Distress .(tolerance (DBT
L2 Capture,Motivated Reasoning : L2 distorts reality to justify L1 impulses.,"Rigid certainty; elaborate ""logical"" defenses for mal-adaptive choices.",CBT; Socratic questioning; Targeted .journaling

L3 Depletion,Metabolic exhaustion; PFC lacks fuel for the Explore-Exploit Gap.,Apathy; decision fatigue; loss of inhibitory control .,"L0 Restoration (Sleep, Nutrition, Load ".(reduction
L3 Confabulation,Illusory Agency : L3 narrates a choice already made by L1/L2.,"Feeling in control while repeating destructive, automated patterns.",ACT; Cognitive defusion; .Mindfulness
Autopilot,Low-fidelity active inference : L1+L2 drive behavior without L3.,"Numbness; ""going .through the motions""; passenger mode.",Intentional disruptions; Crisis; Existential inquiry
L0 Collapse,"Sequential failure; L3 fails first, then L2, then L1 amplifies.",Burnout; nervous .breakdown; total hardware shutdown.,Psychiatry; Medical safety; Intensive sleep protocols

4. Clinical Applications: Precision Triage and the "Unified API"

The Failure State taxonomy serves as a meta-diagnostic tool—a **Unified API**—to align therapeutic selection with the specific layer of malfunction. Traditional interventions often fail .because they are misaligned with the patient's current failure state

The Problem of Secondary Shame

A critical clinical finding of the L-Stack is that attempting L2 (Cognitive) work during an **L0 Collapse** or **L1 Override** is not merely ineffective, but actively harmful. It induces **Secondary Shame** by demanding the Pilot (L3) perform executive tasks when the biological hardware is literally "throttled" and offline. This creates a feedback loop of .perceived failure that further degrades the L0 substrate

Precision Triage Protocol

1. **Stabilize L0:** Restore biological integrity (Sleep, Nutrition, Allostatic load reduction).
2. **Regulate L1:** Utilize somatic tools and exhalatory modulation to dampen the .amplitude of prediction errors
3. **Engage L2/L3:** Once the substrate is stable and signals are regulated, initiate .(cognitive restructuring (CBT) or executive "Explore-Exploit" training (ACT

5. Cybernetic Convergence: AI Safety and the Artificial L3 Pause

The L-Stack provides a direct bridge between human cognitive architecture and the .emerging levels of Agentic AI autonomy

Human Architecture vs. AI Autonomy

Hierarchical Level,AI/Agentic Level (SAE),Core Mechanical Function
L0,Manual / No Autonomy,Foundational Hardware Constraints
L1/L2,Stateless Reactive Autonomy,Script Execution / Pattern Matching
L3,Conditional Orchestration,Error Correction / Evaluation Gap
L4,High Autonomy,Proactive Environmental Interaction
L5,Generative Autonomy,System Governance / Rule Definition

The Artificial L3 Pause

For highly autonomous AI (L4+), this model argues for a mandatory **Artificial L3 Pause**. This is a computationally expensive verification phase where the system must temporarily suspend its immediate execution (exploitation) to evaluate ethical alignment and truthfulness (exploration). This mechanism is designed to prevent **Artificial L2 Capture**

(hallucinations/rationalizing errors) and **Artificial L1 Override** (unbounded algorithmic execution), ensuring the agent remains aligned with human intent.

6. Conclusion: The Future of Cognitive Agency

The L-Stack represents a transformative framework for clinical, organizational, and technological domains. By recognizing that human agency resides solely within the "Explore-Exploit Gap," we can protect this capacity from L5 systems—specifically attention-economy algorithms—designed to erode L3 executive function in favor of L1 reactivity.This framework is not a theoretical abstraction; it is a survival-driven synthesis. It was stress-tested under conditions of extreme environmental volatility, including missile attacks, systemic institutional failure, and acute personal grief. These "black swan" events validated the L0-L5 priority order by demonstrating that when the hardware (L0) is under fire, the "Pilot" (L3) is the first system to go offline.Whether we are treating a patient, leading an organization, or designing an AI, our primary objective must be the protection and strengthening of the L3 executive gap. Genuine freedom is not found in the scripts we follow, but in the metabolic capacity to stop and evaluate them.

Tal Nagar - All Rights Reserved 2026 ©